Paternity analysis in Excel

Margarida Rocheta
Centro de Botânica Aplicada à Agricultura, Secção de Genética,
Instituto Superior de Agronomia (ISA), TU Lisbon, Portugal
rocheta@isa.utl.pt

F. Miguel Dionísio*, Luís Fonseca, Ana M. Pires**
Departamento de Matemática, Instituto Superior Técnico (IST), TU Lisbon, Portugal
{fmd,apires}@math.ist.utl.pt

*also at SQIG, Instituto de Telecomunicações (IT), Lisbon, Portugal
**also at the Center for Mathematics and its Applications, Lisbon, Portugal

Abstract

Paternity analysis using microsatellite information is a well studied subject. These markers are ideal for parentage studies and fingerprinting, due to their high discrimination power. This type of data is used to assign paternity, to compute the average selfing and outcrossing rates and to estimate the biparental inbreeding. There are several public domain programs that compute all this information from data. Most of the time, it is necessary to export data to some sort of format, feed it to the program and import the output to an Excel book for further processing. In this article we briefly describe a program referred from now on as **PAE** (**P**aternity **A**nalysis in **E**xcel), developed at IST and IBET (see the acknowledgments) that computes paternity candidates from data, and other information, from within Excel. In practice this means that the end user provides the data in an Excel sheet and, by pressing an appropriate button, obtains the results in another Excel sheet. For convenience **PAE** is divided into two modules. The first one is a filtering module that selects data from the sequencer and reorganizes it in a format appropriate to process paternity analysis, assuming certain conventions for the names of parents and offspring from the sequencer. The second module carries out the paternity analysis assuming that one parent is known. Both modules are written in Excel-VBA and can be obtained at the address `www.math.ist.utl.pt/~fmd/pa/pa.zip`. They are free for non-commercial purposes and have been tested with different data and against different software (Cervus, FaMoz, and MLTR).

1 Introduction

Microsatellite information is obtained experimentally using single sequence repeat (SSR) primers that are analyzed in a sequencer. Post-processing of these data for parentage studies include

1. filtering, i.e. selection of relevant data ignoring values provided by the sequencer that are outside a (user provided) range;

2. computation of diversity parameters such as the observed number of alleles (A_o), the observed heterozygosity (H_o), the expected heterozygosity (H_e) ([1]), and the fixation index ($F = 1 - H_o/H_e$) ([2]), computation of polymorphic information

content (PIC) ([3]), exclusion probabilities ([4]) and testing of Hardy-Weinberg equilibrium (HWE) departures for each locus;

3. identification of the most likely father(s) using the likelihood approach of Meagher and Thompson ([5], see also [6]).

There are different public domain programs, like FaMoz ([7]), Cervus 2.0 ([6]) and MLTR 2.2 ([8]), that do some of the computations required for paternity analysis but they are not necessarily user-friendly. For a sporadic user it takes a lot of time and effort to understand how the input and the output works. Our purpose was to create a simple way to analyze paternity data without the need to use different kinds of software. **PAE** was designed to treat data in an Excel sheet, independently of the sequencer used. It is able to filter the data and to compute all the necessary values for the analysis using the raw data from the sequencer avoiding thus a painful formatting process required by other programs. This is in our opinion a great advantage. It is worth mentioning that **PAE** was tested on real data and that the results were consistent with those obtained with the software previously cited.

We describe the computations and some technical details in the next section. The program and user interface are introduced in Section 3. Section 4 concludes the article.

2 Computational methods and theory

While a detailed discussion of the statistical and mathematical methods used for paternity analysis is beyond the scope of this paper, a brief description of the mathematical expressions implemented in **PAE** is given here for completeness. For a more comprehensive introduction to the subject we recommend the review by Jones and Ardren ([9]).

Let n be the number of genotypes, $2n$ the total number of alleles per locus (assuming diploid organisms) and L the number of loci.

For a given locus j ($j = 1, \ldots, L$) with $k(j)$ alleles let $n_i(j)$ denote the observed count of the i^{th} allele, A_i, $i = 1, \ldots, k(j)$, and $n_{ii}(j)$ the observed count of homozygotics with the i^{th} allele (genotype A_iA_i). The observed heterozygosity is

$$H_o(j) = 1 - \sum_{i=1}^{k(j)} \frac{n_{ii}(j)}{n}$$

and the (estimated) expected heterozygosity is

$$H_e(j) = 1 - \sum_{i=1}^{k(j)} \tilde{p}_i^2(j), \tag{1}$$

where $\tilde{p}_i(j) = n_i(j)/(2n)$ is the observed frequency of the i^{th} allele (note that $\sum_{i=1}^{k} n_i(j) = 2n$ and hence $\sum_{i=1}^{k} \tilde{p}_i(j) = 1$). The fixation index, $F(j) = 1 - H_o(j)/H_e(j)$, compares the observed and the estimated expected heterozygosity, which is derived under HWE. The (estimated) polymorphic information content (PIC) is given by

$$PIC(j) = H_e(j) - \sum_{i=1}^{k(j)-1} \sum_{l=i+1}^{k(j)} 2\,\tilde{p}_i^2(j)\,\tilde{p}_l^2(j) \tag{2}$$

and is often used to measure the informativeness of a genetic marker for linkage studies, irrespectively of the mode of inheritance.

The usefulness of a co-dominant marker for parentage testing is determined by the probability of that marker implying an exclusion. For one known parent and one offspring the (estimated) probability of excluding their relationship based on the genotypes at the j^{th} locus can be computed as

$$P_e(j) = 1 - 4\sum_{i=1}^{k(j)} \tilde{p}_i^2(j) + 2\left(\sum_{i=1}^{k(j)} \tilde{p}_i^2(j)\right)^2 + 4\sum_{i=1}^{k(j)} \tilde{p}_i^3(j) - 3\sum_{i=1}^{k(j)} \tilde{p}_i^4(j). \tag{3}$$

Combining expressions (1), (2) and (3) over the L unlinked loci leads to:

- mean (estimated) expected heterozygosity

$$H_e = \frac{1}{L}\sum_{j=1}^{L} H_e(j) = 1 - \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{k(j)} \tilde{p}_i^2(j),$$

- mean (estimated) *PIC*

$$PIC = \frac{1}{L}\sum_{j=1}^{L} PIC(j),$$

- and global (estimated) exclusion probability

$$P_e = 1 - \prod_{j=1}^{L}(1 - P_e(j)).$$

Note that the theoretical counterpart of formulae (1), (2) and (3) are similar to those but with $\tilde{p}_i(j)$ replaced by $p_i(j)$, the theoretical, usually unknown, probability of the i^{th} allele for the j^{th} locus.

The chi-square test is used to test Hardy-Weinberg equilibrium (HWE) for each locus ($j = 1, \ldots, L$), that is, to test the null hypothesis: (i) the probability of a homozygotic genotype A_uA_u is the product of the probabilities of the corresponding alleles ($P_{uu}(j) = p_u^2(j)$), and (ii) the probability of a heterozygotic genotype A_uA_v is twice the product of the probabilities of the corresponding alleles ($P_{uv}(j) = 2\,p_u(j)p_v(j)$). The observed value of the chi-square statistic is given by

$$X_0^2(j) = \sum_{u=1}^{k^*(j)}\sum_{v=u}^{k^*(j)} \frac{(|n_{uv}(j) - e_{uv}(j)| - 0.5)^2}{e_{uv}(j)}, \tag{4}$$

where $n_{uv}(j)$ denotes as before the observed count of the A_uA_v genotype and $e_{uv}(j)$ is the corresponding estimated expected count under the null hypothesis, that is,

$$e_{uv}(j) = \begin{cases} n\,\tilde{p}_u^2(j), & u = v \\ 2n\,\tilde{p}_u(j)\tilde{p}_v(j), & u \neq v \end{cases}$$

$k^*(j)$ in (4) denotes the reduced number of alleles at locus j after considering as one fictitious "allele" the set of all alleles not verifying $n\tilde{p}_u^2(j) > 5$. The number of degrees of freedom of the chi-square statistic is $df(j) = k^*(j)(k^*(j) - 1)/2$. Finally the p-value of the test is $P(Q > X_0^2(j))$, with $Q \sim \chi^2_{df(j)}$. The observed value is usually regarded as non-significant, that is giving evidence that the j^{th} locus is under HWE, if the associated p-value is greater than 5%.

The expressions used to identify the most likely father(s) (recall that the mother is assumed as known) are taken from [6]. Those are based on the likelihood ratio test for the hypotheses H_0: "the alleged father is the true father" against H_1: "the true father is an individual selected randomly from the population". If g_m, g_a and g_o denote, respectively, the genotype of the mother, the alleged father and the offspring, at a given locus, j, the likelihood ratio (also known as Paternity Index, PI) is given by

$$PI(j) = L(H_0, H_1 | g_m, g_a, g_o) = \frac{P(g_o|g_m, g_a)}{P(g_o|g_m)}, \tag{5}$$

where $P(g_o|g_m, g_a)$ ($P(g_o|g_m)$) denotes the probability of the offspring genotype given the mother and the alleged father genotype (given only the mother genotype). Table 1 lists the possible values of (5).

Table 1: Likelihood ratios for the compatible genotypes of offspring (g_o), alleged father (g_a) and mother (g_m). U and V are generic alleles, X represents any allele other than U and Y represents any allele other U and V. $\tilde{p}_u$ and $\tilde{p}_v$ denote the observed frequencies of alleles U and V at the locus under consideration, respectively (the explicit reference to the locus, (j), has been omitted for simplicity).

g_o	g_a	g_m	$P(g_o\|g_m, g_a)$	$P(g_o\|g_m)$	$L(H_0, H_1\|g_m, g_a, g_o)$
UU	UU	UU	1	$\tilde{p}_u$	$1/\tilde{p}_u$
UU	UX	UU	1/2	$\tilde{p}_u$	$1/(2\tilde{p}_u)$
UU	UU	UX	1/2	$\tilde{p}_u/2$	$1/\tilde{p}_u$
UU	UX	UX	1/4	$\tilde{p}_u/2$	$1/(2\tilde{p}_u)$
UV	UU	VV	1	$\tilde{p}_u$	$1/\tilde{p}_u$
UV	UX	VV	1/2	$\tilde{p}_u$	$1/(2\tilde{p}_u)$
UV	UU	VY	1/2	$\tilde{p}_u/2$	$1/\tilde{p}_u$
UV	UX	VY	1/4	$\tilde{p}_u/2$	$1/(2\tilde{p}_u)$
UV	UU	UV	1/2	$(\tilde{p}_u + \tilde{p}_v)/2$	$1/(\tilde{p}_u + \tilde{p}_v)$
UV	UY	UV	1/4	$(\tilde{p}_u + \tilde{p}_v)/2$	$1/[2(\tilde{p}_u + \tilde{p}_v)]$
UV	UV	UV	1/2	$(\tilde{p}_u + \tilde{p}_v)/2$	$1/(\tilde{p}_u + \tilde{p}_v)$

Then, for each alleged father, the paternity indexes for each locus are multiplied, assuming again unlinked loci, and the natural logarithm taken. This leads to what is usually designated the LOD score,

$$LOD(g_o, g_a) = \prod_{j=1}^{L} PI(j).$$

Finally, for each offspring, the positive LOD scores are sorted and the most likely father is found (a LOD score equal to zero means that the alleged father is as likely of being the father as any other randomly selected candidate, while a positive one means that

the alleged father is more likely the father than a randomly selected candidate). The difference between the two most likely fathers, Δ, is also computed.

The program computing all the values just described has been written in Visual basic (VBA), an object oriented language that supports programming in Excel. In particular it has predefined classes corresponding to Excel concepts such as rows and columns. The previous "formulas" are computed by storing the data in "arrays", processing those in auxiliary arrays. The code is available in the provided VBA modules (see installation instructions at `www.math.ist.utl.pt/~fmd/pa/pa.zip`).

The program has been used for paternity analysis of *Pinus pinaster* at ITQB-IBET (www.itqb.unl.pt), and its results fully agree with other software namely FaMoz ([7]), Cervus 2.0 ([6]) and MLTR 2.2 ([8]). The data on which **PAE** is illustrated in the next section contains the genotypes on three microsatellite markers for 60 parental trees and 206 offspring.

3 Program and user interface

The two modules of **PAE** are used from within an Excel sheet[1]. The Excel sheet should be the subset of the sequencer output including only the columns with the name of individuals, estimated fragment size, peak area and peak height (see Figure 1).

Microsoft Excel - exampleSequencer.xls

File Edit View Insert Format Tools Data Window Help

A1 = RN

	A	B	C	D
1	RN	est frag size (nt)	pk area (rfuxmm)	pk height (rfu)
2	e61XII/T-1-94.E02_06101516J5	359,94	4205	7868
3	e61XII/T-1-94.E02_06101516J5	380,25	3937	7354
4	e61XII/T-1-94.E02_06101516J5	54,59	4438	7290
5	e61XII/T-1-94.E02_06101516J5	55,57	45688	67899
6	e61XII/T-1-94.E02_06101516J5	146,32	4953	8744
7	e61XII/T-1-94.E02_06101516J5	148,41	9639	17505
8	61I/E-94.H08_06101516J6	55,48	5721	8948
9	61I/E-94.H08_06101516J6	144,58	3846	6774
10	61I/E-94.H08_06101516J6	146,69	11987	21504
11	61I/E-94.H08_06101516J6	148,78	26237	45324
12	61I/E-94.H08_06101516J6	161,01	3035	5397
13	61I/E-94.H08_06101516J6	171,37	1910	3682

Figure 1: Raw data from the sequencer (excerpt), that will be used as input for the filtering process. Only the shown columns are needed. Note that names of individuals must follow conventions described in the text, in particular, the identification of parents must be included in those of offspring.

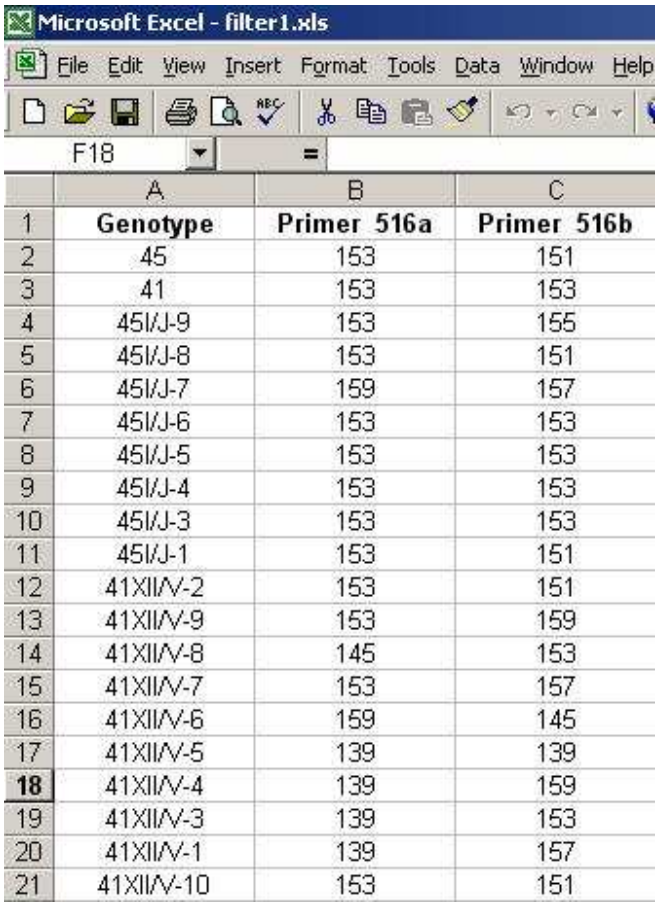
The convention for the names of individuals is the following: candidate parents have a name that begins with a number identifying them. Offspring names must begin with an **e** (from embryo) followed by the number identifying the known parent. In the example, **e61** refers to offspring of (mother) **61**. The names may contain other data (ex. **XII/T**) that identifies for instance the geographic position of individuals and is irrelevant for

[1]Installation instructions are included in `www.math.ist.utl.pt/~fmd/pa/pa.zip`.

the program. Different offspring of the same parent are identified by a number as in **e61XII/T-1-94** or **e61XX/B-2-94**. The primer must also occur in the name both of parents and offspring following an hyphen. In these examples the primer is 94.

PAE assumes that all parents occur together at the beginning of the Excel sheet, before offspring. For this reason it may be necessary to sort the data (by the first column).

Activation (by clicking) on the button *Filter and Prepare* produces another Excel sheet with the filtered data. The user is asked interactively to provide the lower and upper bound for the values of the alleles (estimated fragment size - base pairs). It may be necessary to repeat filtering with different bounds until the filtered data contains exactly one line per individual. The filter program assumes that the data come from diploid individuals. The output of filtering, i.e. the peaks that best represent the alleles are contained in a new sheet called Selection. Another sheet, named New Disposition contains the same data but in a form appropriate to paternity analysis (see Figure 2).



	A	B	C
1	**Genotype**	**Primer 516a**	**Primer 516b**
2	45	153	151
3	41	153	153
4	45I/J-9	153	155
5	45I/J-8	153	151
6	45I/J-7	159	157
7	45I/J-6	153	153
8	45I/J-5	153	153
9	45I/J-4	153	153
10	45I/J-3	153	153
11	45I/J-1	153	151
12	41XII/V-2	153	151
13	41XII/V-9	153	159
14	41XII/V-8	145	153
15	41XII/V-7	153	157
16	41XII/V-6	159	145
17	41XII/V-5	139	139
18	41XII/V-4	139	159
19	41XII/V-3	139	153
20	41XII/V-1	139	157
21	41XII/V-10	153	151

Figure 2: Filtered data (excerpt) with different disposition: for each individual (genotype) the filtered values of the alleles corresponding to each primer are displayed. Note that individuals (genotype) have been renamed (see text) and that the lines referring to parents precede those referring to offspring.

Recall that the primer is the string that follows the hyphen in the name of parents and the string following the second hyphen in the names of offspring. It is mandatory to follow this convention for New Disposition to be built correctly. Note that the names of

offspring are no longer prefixed by **e**. In fact the conventions for the names of parents and offspring are needed only for filtering. **PAE** assumes that all parents occur before offspring and that the names of offspring begin with the number identifying the known parent.

After filtering, the computations for paternity analysis are started by clicking on the button *Paternity Analysis*, in the sheet New Disposition. The user is then asked to provide the number of parents and offspring involved. After a short time another Excel sheet is produced containing the analysis of the data. This includes a first table containing, for each locus, the number of alleles, the number of individuals, the number of heterozygotics, the number of homozygotics, the frequency of heterozygotics observed, the polymorphic information content (PIC) and the exclusion probability.

For each locus the information associated with its alleles is displayed in another table containing the alleles, the count of their occurrence in the locus, number of heterozygotics, the number of homozygotics and the frequency of alleles. Moreover, the PIC and Hardy-Weinberg equilibrium test values (when possible) are also shown. Finally, paternity analysis is presented in another table giving for each offspring (6^{th} column) the two most likely fathers (see Figure 3).

LN	Nr mismatch Father|Offspring	[illegible]	Candidate Father 1	Candidate Father 2	Offspring	Delta
1,910408296	1	1	11	16	23IH-1	0,554163757
3,525644743	0	0	45	40	23IH-2	0,194800993
2,735418211	0	1	23	34	23IH-3	1,058739376
2,039551484	0	1	23	48	23IH-4	0,756248023
3,735061053	0	0	23	33	23IH-5	3,043846646
3,525644743	0	0	45	40	23IH-6	0,194800993
2,924687686	0	1	23	29	23IH-7	2,378482089
2,195129699	1	1	29	24	23IH-8	0,315471051
2,697287324	0	0	58	63	23IH-9	1,25221549
2,69742588	0	1	23	45	23IH-10	1,911499321
2,997503594	0	0	33	24	33IB-1	1,16769773
2,785833315	0	0	8	18	33IB-2	0,630913787
3,328961904	0	0	80	18	33IB-3	0,646228033
2,746712491	0	0	61	53	33IB-4	0,102033105
2,558117402	0	0	35	61	33IB-5	0,004284254
1,159629074	1	1	46	47	33IB-6	0,214652644

Figure 3: Most likely fathers (excerpt). Each line contains information referring to one offspring. Refer to the text for a description of the columns in the table. Colours code the distance between the two most likely fathers. Sepia means that there is clearly one most probable father and green that there is a good candidate. Lines not coloured mean that the analysis cannot clearly assign a candidate father to that offspring.

The first column displays the *LOD* score for the most likely father, the second the number of father/offspring mismatches, the third the number of father/offspring mismatches given the mother, the fourth and fifth the identification of the two most likely fathers, the sixth the name of the offspring these refer to and finally the seventh column displays the distance between the two most likely fathers (Δ). The colors code the distance between the two most likely fathers. Sepia means that $\Delta > 1.2$ and there is clearly only one most probable father. Green means $0.6 < \Delta < 1.2$ and that there is a considerable distance between the most probable father and the second candidate father. Lines that are not colored mean that $\Delta < 0.6$ and that it is not clear which of the two is the father. The number of offspring per father is presented in an additional table.

4 Conclusion and availability

We have briefly described programs to filter data from a sequencer and perform paternity analysis, from within Excel, available at `www.math.ist.utl.pt/~fmd/pa/pa.zip`. We expect them to be useful and that they will save time to researchers in this field. We do not plan to add new features to the code but please feel free to edit and change the code to suit your needs.

Acknowledgments

This work started at Instituto de Biologia Experimental e Tecnológica (ITQB-IBET, `www.itqb.unl.pt`). It was partially supported by FCT and EU FEDER, namely via CEMAT POCTI and CLC POCTI (Research Unit 1-601).

References

[1] M. Nei, Molecular Evolutionary Genetics, Columbia University Press, New York NY, 1987.

[2] B.S. Weir and C.C. Cockerham, Estimating F-statistics for the analysis of population structure, Evolution 38 (1984) 1358-1370.

[3] D. Botstein, R.L. White, K. Skolnick and R.W. Davis, Construction of a genetic linkage map in man using restriction fragment length polymorphism, American Journal of Human Genetics 32 (1980) 314-331.

[4] A. Jamieson and St.C.S. Taylor, Comparisons of three probability formulae for parentage exclusion, Animal Genetics 28 (1997) 397-400.

[5] T.R. Meagher and E. Thompson, The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction, Theoretical Population Biology 29 (1986) 87-106.

[6] T.C. Marshall, J. Slate, L. Kruuk and J.M. Pemberton, Statistical confidence for likelihood-based paternity inference in natural populations, Molecular Ecology 7 (1998) 639-655.

[7] S. Gerber, P. Chabrier and A. Kremer, FaMoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers, Molecular Ecology Notes 3 (2003) 479-481.

[8] K. Ritland, Extensions of models for the estimation of mating systems using n independent loci, Heredity 88 (2002) 221-228.

[9] A.G. Jones and W.R. Ardren, Methods of parentage analysis in natural populations, Molecular Ecology 12 (2003) 2511-2523.